High Frequency Data

Chapter 3

Financial Econometrics

Michael Hauser

WS17/18

Content

- Data frequencies and volatility measures
- Distribution of returns:
  monthly, daily, intraday data
- Variance ratio test for WN,
  small sample distribution via resampling
- Intraday microstructure:
  bid-ask bounce, asynchrone trading, parent and child orders
- Measures for volatilities:
  monthly, daily and variants,
  intraday: realized volatility

Stock returns, aggregation

Stock return data

- Monthly data.
  Returns are defined as the return between last price of the previous month and the last price of the current month.
  Since the number of days vary in a deterministic way across months, returns will tend to show a seasonal pattern, unless the number of trading days is corrected for.
- Daily data.
  Several prices are recorded:
  Open, O, Close, C, High, H, Low, L.
  Returns are commonly defined as the difference of the Close of the current day and the Close of the previous day.
  Weekends are generally ignored. Bank holidays are filled with the price of the previous day.

Stock return data

- Intraday data.
  - Intraday data are collected at a frequency smaller than 24 hours, going down to tick data Bid, Ask and Volume.
  - In case of heavy trading even trades with different Bids/Asks and Prices at the same tick are observed.
  - Usually all data are discrete:
  prices jump for discrete amounts like $1/4$ Dollar;
  there are minimal trade sizes;
  there are always a finite number of trades per day.
  - Tick data have generally to be controlled for data errors and institutional idiosyncracies (e.g. finding of the opening price of a stock exchange) before analysis.

[Plots of paths and distributions of returns: monthly, daily, intraday.]

Aggregation of returns

We start with daily data. Daily returns, r_t, are defined as the difference of the log prices.

$$r_t = \log(P_t) - \log(P_{t-1}) = p_t - p_{t-1}$$

Then the *monthly* return, $r_t^{(m)}$, is the sum of all *daily* returns of this month with m days:

$$P_{t-m} \to P_{t-m+1}, \quad \ldots \quad , P_{t-1} \to P_t$$

$$r_t^{(m)} = \log(P_t) - \log(P_{t-m}) =$$

$$= p_t - p_{t-m} = [p_t - p_{t-1}] + [p_{t-1} - p_{t-2}] + \ldots + [p_{t-m+1} - p_{t-m}] =$$

$$= \sum_{i=1}^{m} r_{t-i+1}$$

This is a nice property of log-returns.

Aggregation of returns: variances

$$\mathrm{E}(r_t) = \mu, \qquad \mathrm{E}(r_t^{(m)}) = m\,\mu$$

The variances of the daily and monthly returns are with $r_t = \mu + \epsilon_t$

$$\mathrm{V}(r_t) = \mathrm{V}(\epsilon_t) = \sigma^2$$

$$\mathrm{V}(r_t^{(m)}) = \mathrm{E}\left(\sum_{i=1}^{m} r_{t-i+1} - m\,\mu\right)^2 = \mathrm{E}\left(\sum_{i=1}^{m}(r_{t-i+1} - \mu)\right)^2 = \mathrm{E}\left(\sum_{i=1}^{m}\epsilon_{t-i+1}\right)^2$$

Empirically, the variance of the monthly return is in general *not* equal to the sum of the daily variances.

Only in case of $r_t = \mu + \epsilon_t$, with ϵ_t WN,

$$\mathrm{V}(r_t^{(m)}) = \mathrm{E}\left(\sum_{i=1}^{m}\epsilon_{t-i+1}\right)^2 \overset{(!)}{=} \sum_{i=1}^{m}\mathrm{E}(\epsilon_{t-i+1})^2 = m\,\sigma^2 = m\,\mathrm{V}(r_t)$$

holds.

Aggregation of returns: variances

For 2 rvs X and Y $\quad \mathrm{V}(X+Y) = \mathrm{V}(X) + 2\mathrm{Cov}(X,Y) + \mathrm{V}(Y)$.
Whether $\mathrm{V}(X+Y)$ is larger or smaller than $\mathrm{V}(X) + \mathrm{V}(Y)$ depends on the sign of the covariance.

If positive (negative) autocorr dominates in ϵ_t, than the monthly variance per day (i.e. divided by m) is larger (smaller) than the daily variance.

It may happen that autocorrelations cancel out. (E.g. postive autocorr at lag 3 cancels with negative autocorr at lag 5.)

Variance ratio test, VR-test

The **Variance ratio test** is a test for WN. It compares the variance of daily returns with the variance of the data with monthly frequency.
Say we observe M months of equal length m, $T = m\,M$.
σ_A^2 is the variance of the daily returns,
σ_B^2 is the variance of the monthly returns per day.
$\widehat{\mu} = (1/T)\sum_{t=1}^{T} r_t$

$$\sigma_A^2 = \frac{1}{T}\sum_{t=1}^{T}(r_t - \widehat{\mu})^2, \qquad \sigma_B^2(m) = \left[\frac{1}{M}\sum_{\tau=1}^{M}(r_\tau^{(m)} - m\,\widehat{\mu})^2\right]/m$$

$$VR(m) = \frac{\sigma_B^2(m)}{\sigma_A^2} - 1$$

Then under the null hypothesis of WN population values of σ_A^2 and σ_B^2 are equal.

$$\sqrt{T}\,VR(m) \overset{a}{\sim} \mathrm{N}(0, 2(m-1))$$

Variance ratio test, VR-test

This is a *non overlapping* (months) version of the VR-test.
The *overlapping* one has a higher power as more observations for the ”monthly” data are available.

The small sample distribution of the test statistic may be rather different especially in case of heteroscedastic returns. Therefore simulation of the distribution via *resampling* is recommended to obtain critical values, or to refer to special tabulated values.

Resampling destroys possible time dependencies, but keeps the univariate distributional properties, like fat tails.

Volatilities for daily returns

Measuring volatilities for daily data

Models for conditional heteroscedasticity like GARCH use often daily returns. They interpret the squared error of r_t *after taking out* possible linear effects, $(r_t - \mathrm{E}(r_t|I_{t-1}))^2$, as local variance. I_{t-1} stands for the information available at $(t-1)$.
If information about the within day behavior is available, this estimate can be improved.

If we assume that the underlying process is a continuous Brownian motion without drift, then using the information of the daily High and Low instead can be backed up theoretically.

A simple measure is

the *range* of the price, $H_t - L_t$.

It will, however, underestimate the true range if the tick size of the stock is large or the stock is traded not frequently.

Estimates of the daily variance

Other simple estimates (in the notation of Tsay) are

$$\widehat{\sigma}_0^2 = (C_t - C_{t-1})^2$$

$$\widehat{\sigma}_2^2 = 0.3607(H_t - L_t)^2$$

$$\widehat{\sigma}_5^2 = 0.5(H_t - L_t)^2 - 0.386(C_t - O_t)^2$$

The first, $\widehat{\sigma}_0^2$, is the **squared daily return**. *Relative efficiencies* of these measures are

$$\mathrm{V}(\widehat{\sigma}_0^2)/\mathrm{V}(\widehat{\sigma}_j^2), \qquad j = 2, 5$$

j	2	5
rel. eff.	5.2	7.4

I.e. under Brwonian motion without drift the variance of $\widehat{\sigma}_0^2$ is 5.2 times larger than the variance of $\widehat{\sigma}_2^2$.

Estimates of the daily variance

σ_0^2 also covers the overnight period, while the other measures do not. So the difference between them depends on the assumption about the behavior during night: no new information, Brownian motion, or jumps.

Autocorrelated intraday returns: Bid-ask bounce

Intraday data: Bid-Ask spread

The difference between the ask price and the bid price, $S = P_a - P_b$, is the *bid-ask spread*. It is seen as a market friction necessary to make the market work.

The bid-ask spread introduces a *negative* lag-1 serial correlation in the observed asset return, even if the underlying process is WN. This is referred to as the **bid-ask bounce**.

Say, realized prices, P_t, are based on the underlying price, P_t^*.

$$P_t = P_t^* + I_t\,(S/2)$$

- P_t^* is the true, but unobserved price. P_t^* is a RW.
- I_t is an independent rv, which takes the values ± 1 with prob 0.5 (and so is stationary). It indicates whether the *market maker* sells $(+1)$ or buys (-1) in a trade.

Bid-ask bounce

- t is a time point, where a trade takes place.
- Properties of P_t^*:
  We assume P_t^* is a RW (without drift) with ϵ_t WN.
  I_t and ϵ_t are independent.

  $P_t^* = P_{t-1}^* + \epsilon_t,$ $\Delta P_t^* = \epsilon_t$ is WN.
- Properties of I_t:
  $\mathrm{E}(I_t) = 0$ and $\mathrm{V}(I_t) = 1.$
  $\mathrm{E}(I_t - I_{t-1}) = \mathrm{E}(\Delta I_t) = 0$ and $\mathrm{V}(I_t - I_{t-1}) = \mathrm{V}(\Delta I_t) = 2.$
  However, ΔI_t is not a WN, but a MA(1): $\mathrm{Cov}(\Delta I_t, \Delta I_{t-1}) = -1.$
  Higher order auto-cov vanish: $\mathrm{Cov}(\Delta I_t, \Delta I_{t-j}) = 0, \ j > 1.$

For simplicity we assume time intervals of fixed length.

Bid-ask bounce: 1-period returns

- Properties of the 1-period return process:

$\Delta P_t = P_t - P_{t-1}$,	$\Delta P_t = \Delta P_t^* + \Delta I_t\,(S/2)$.
Its variance is	$\mathrm{V}(\Delta P_t) = \sigma_\epsilon^2 + 2\,(S^2/4)$.
Its order 1 auto-covariance	$\mathrm{Cov}(\Delta P_t, \Delta P_{t-1}) = -S^2/4$.
Its order 1 auto-correlation is	$[-S^2/4]/[\sigma_\epsilon^2 + S^2/2]$.
Higher order auto-cov vanish	$\mathrm{Cov}(\Delta P_t, \Delta P_{t-j}) = 0, \;\; j > 1$.

So, the observed 1-period return process is autocorrelated with order 1 (only). This is similar to the problem of overdifferencing, as the difference of a WN is a MA(1).

Bid-ask bounce: k-period returns, $\sigma_\epsilon^2(k) = k\,\sigma_\epsilon^2(1)$

The k-period *non overlapping* return process, $\Delta_k P_\tau$, is

$$\Delta_k P_\tau = P_\tau - P_{\tau-k} = \Delta_k P_\tau^* + \Delta_k I_\tau\,(S/2)$$

$\tau = n\,k$, $n = 1, 2, \ldots$: $\quad \tau = k, 2\,k, 3\,k, \ldots$.
We investigate its autocorrelation structure.

- $\Delta_k P_\tau^*$: $\qquad \Delta_k P_\tau^* = \sum_{t=\tau-k+1}^{\tau} \epsilon_t$

  $\mathrm{E}(\Delta_k P_\tau^*) = 0, \quad \mathrm{V}(\Delta_k P_\tau^*) = k\,\sigma_\epsilon^2.$

  All auto-covariances of order greater zero of $\Delta_k P_\tau^*$ vanish, since the ϵ_t's do not overlap.

- $\Delta_k I_\tau = I_\tau - I_{\tau-k}$:

  $\mathrm{E}(\Delta_k I_\tau) = 0, \quad \mathrm{V}(\Delta_k I_\tau) = 2,$

  $\mathrm{Cov}(\Delta_k I_\tau, \Delta_k I_{\tau-k}) = \mathrm{E}(I_\tau - I_{\tau-k})(I_{\tau-k} - I_{\tau-2k}) = -1$

  Higher order covariances are zero.

Bid-ask bounce: k-period returns

- $\Delta_k P_\tau$:

  $$\mathrm{E}(\Delta_k P_\tau) = 0, \qquad \mathrm{V}(\Delta_k P_\tau) = k\,\sigma_\epsilon^2 + 2\,(S^2/4).$$

  $$\mathrm{Cov}(\Delta_k P_\tau, \Delta_k P_{\tau-k}) =$$
  $$= \mathrm{Cov}(\Delta_k P^*_\tau, \Delta_k P^*_{\tau-k}) + \mathrm{Cov}(\Delta_k I_\tau, \Delta_k I_{\tau-k}) S^2/4 = -S^2/4$$

  1st order auto-corr in τ-time is

  $$\mathrm{Corr}(\Delta_k P_\tau, \Delta_k P_{\tau-k}) = [-S^2/4]/[k\sigma_\epsilon^2 + (S^2/2)]$$

  Higher order auto-covariances are zero:

  $$\mathrm{Cov}(\Delta_k P_\tau, \Delta_k P_{\tau-jk}) = 0, \quad j > 1.$$

Here we also find order 1 auto-corr in the returns. But with S small and k large it is negligible.

The k-period return is approximately WN for large k.

Autocorrelated intraday returns: Asynchronous trading

Asynchronous trading

Asynchronous or nonsynchronous trading implies that the underlying (continuous) return process is observed at irregular times. The consequences are serially correlated observed returns, even if the underlying prices are a pure RW *with drift* μ (the returns are iid).

We think of equidistant time points at which a trade *can* occur: $t = 1, 2, 3, \ldots$.

We distinguish periods where a trade occurs and periods with no trade.

When we observe a return, r_t^o, *after k periods with no trade*, it is the sum of all time step underlying/unobserved returns since the last trade.

$$r_t^o = \sum_{i=0}^{k} r_{t-i} \quad r_j \text{ unobserved}$$

If there is no trade in t, we refer to the last observed price, so that $r_t^o = 0$.

Asynchronous trading

- We assume the underlying return process r_t is iid and $\sim(\mu, \sigma^2)$.
- We distinguish
  periods of *no* trade with prob π and
  periods with a trade with prob $(1-\pi)$.
- k is the number of *'no trade periods'* between 2 trades, $k \geq 0$.
- In the periods with a trade: The probs for $k = 0, 1, 2, \ldots$ are assigned according to the geometric distribution with density $(1-\pi)\pi^k = P(X = k)$.

k	NO	0	1	...	k	...
prob	π	$(1-\pi)^2$	$(1-\pi)^2\pi$	...	$(1-\pi)^2\pi^k$	...
r_t^o	0	r_t	$r_t + r_{t-1}$	...	$\sum_0^k r_{t-i}$	...

Asynchronous trading

It can be shown after tedious calculations that

$$\mathrm{V}(r_t^o) = \sigma^2 + \frac{2\pi\mu^2}{1-\pi}, \qquad \mathrm{Cov}(r_t^o, r_{t-j}^o) = -\mu^2\pi^j, \quad j \geq 1$$

The auto-covariances decrease with decreasing π, and increasing j.

The observed returns r^o's are independent only if the drift μ is zero, or synchronous trading takes place.

Autocorrelated orders: Parent and child orders

Parent and child orders

Large orders (parent orders) are not activated at once, but are split in small orders (child orders) in order to affect the stock price as little as possible.
By subsequently positing small orders of the same type 'positive autocorrelation' of orders arises.

These parent orders are not observed by the public, but may be detected by algorithms.
Ref: O'Hara(2015)

Realized volatility

Realized volatility

If a stock price $P(t)$ evolves in continuous time according to a geometric Brownian motion without jumps

$$dP(t)/P(t) = \mu(t)\,dt + \sigma(t)\,dz$$

where $\mu(t)$ and $\sigma(t)$ denote the drift and the instantaneous volatility process. The **integrated variance** IV for a (predefined unit) time interval $(t-1, t)$ is

$$IV_t = \int_{t-1}^{t} \sigma^2(s)ds$$

The integrated variance is not directly observable. However, the **realized variance** RV

$$RV_t = \sum_{i=1}^{M} r_{i,t}^2$$

where $M = 1/\Delta$, $p(t) = \log(P(t))$ is.

Realized volatility

$r_{i,t}$ is the Δ**-period intraday return** defined as

$$r_{i,t} = p_{t-1+i\Delta} - p_{t-1+(i-1)\Delta}$$

The *RV* provides a consistent estimator of *IV* as the number of intraday observations increases, or equivalently

$$\Delta \to 0.$$

The resulting error in *RV* may be characterized by asymptotic distribution theory, $\Delta \to 0$, as

$$RV_t = IV_t + \eta_t, \qquad \eta_t \sim N(0, 2\,\Delta\, IQ_t)$$

where $IQ_t = \int_{t-1}^{t} \sigma(s)^4\, ds$ denotes the **integrated quadricity**, *IQ*.
IQ may be consistently estimated by the **realized quadricity**, *RQ*,

$$RQ_t = \frac{M}{3} \sum_{i=1}^{M} r_{it}^4$$

Realized volatility

Because of the microstructure of the market Δ should not be chosen too small. Δ should cover at *least 50 trades*.

Realized volatility is the square root of realized variance, $\sqrt{RV}$.

Example:
We construct for 1-minute intraday prices of Microsoft Corporation, MSFT, a series of 1/2-hour realized variances. As the NASDAQ exchange trading hours are 9:30 to 16:00 13 1/2-hour intervals are available, and assume a unit period length of 1/2 hour. I.e. we get per day 13 intraday volatility estimates (ignoring overnight changes).
If we set e.g. $\Delta = 10\text{min}$, $M = 3$. Since MSFT is a frequently traded stock, we could reduce the step size Δ to 2 - 3 minutes.

Time scales

Empirically 3 different time scales make sense:

- time, as usual
- trading time
- volume time

Forecasting daily intraday volatilities

Several people recommend the **heterogenous autoregression (HAR) model** of Corsi(2009) instead of GARCH or stochastic volatility for forecasting the variance of the returns.

$$RV_t = \beta_0 + \beta_1 RV_{t-1} + \beta_2 RV_{t-1|t-5} + \beta_3 RV_{t-1|t-22} + u_t$$

where

$$RV_{t-1|t-h} = \frac{1}{h} \sum_{i=1}^{h} RV_{t-i}$$

Here the unit time interval is one day. So $RV_{t-1|t-5}$ is the average RV of the last week, $RV_{t-1|t-22}$ is the average RV of the last month.

The model incorporates different types of investors, one with a decision horizon of 1 day, one with 1 week and another 1 month. It approximates conveniently a long-memory dynamics as observed in most realized volatilities.

Estimation via OLS. It can be improved by including GARCH errors and assuming an inverse Gaussian error distribution.

Exercises and References

Exercises

Choose 3.

1G Test stock returns of your choice for WN with the variance ratio test. Use `Ex3_1_month_daily_VRtest_R.txt`.

2G Compare 3 different volatility measures for 2 return series of daily stock prices. Use `Ex3_2_DailyVola_R.txt`.

3 Show that the bid-ask bounce effect does not vanish for (fixed) order k non overlapping returns, if we assume that the underlying return process is WN.

4G Compare the distribution of
(a) daily and intraday log returns
(b) daily and intraday (realized) volatilities
for MSFT. Use `Ex3_4_RealVola_msft_R.txt`.

References

Tsay 3.15, 5
Campbell/Lo/MacKinlay 2
O'Hara(2015), Journal of Financial Economics
Corsi(2009), Journal of Financial Econometrics

Appendix: Asymptotics for Quadratic Variation

Quadratic variation

Let the logarithmic price of a financial asset, denoted by $p_t = \log(P_t)$, follow the stochastic-volatility process

$$p_t = p_0 + \int_0^t \mu(s)ds + \int_0^t \sigma(s)dW(s)$$

where μ and σ are càdlàc. W is a standard Brownian motion and σ is assumed to be independent of W.

The **quadratic variation** process, $[p]_t$, of a sequence of partitions, $\tau_0 = 0 \leq \tau_1 \leq \ldots \leq \tau_n = t$, is defined by

$$[p]_t = \operatorname*{plim}_{n\to\infty} \sum_{j=0}^{n-1} (p_{\tau_{j+1}} - p_{\tau_j})^2.$$

Realized variance, one-day intervals

Focusing on one-day intervals, the continuously compounded within-day returns of day t with sampling frequency M is

$$r_{t,j} = p_{t-1+j/M} - p_{t-1+(j-1)/M}, \qquad j = 1, \ldots, M$$

The **realized variance** over day t is defined by

$$RV_t = \sum_{j=1}^{M} r_{t,j}^2$$

By the theory of quadratic variation of semimartingales, (daily) realized variance converges uniformly in probability to the (daily) quadratic variation process as sampling ferquency of returns approaches infinity, i.e. $M \to \infty$

$$RV_t \ \to \ \int_{t-1}^{t} \sigma^2(s) ds$$

providing a consistent estimate.

Asymptotics for the realized variance

The convergance rate is $\sqrt{M}$, and asymptotic normality holds.

$$\sqrt{M}\,\frac{RV_t - \int_{t-1}^{t} \sigma^2(s)ds}{\sqrt{2\int_{t-1}^{t} \sigma^4(s)ds}} \xrightarrow{d} N(0,1)$$

where $\int_{t-1}^{t} \sigma^4(s)ds$ denotes **integrated quadricity**.

The **fourth-power variation** or **realized quadricity** is a consistent estimator

$$RQ_t = \frac{M}{3}\sum_{j=1}^{M} r_{t,j}^4 \;\rightarrow\; \int_{t-1}^{t} \sigma^4(s)ds$$

for the integrated quadricity.